Tests for Lack-of-Fit of Regression Models

Giovanni C. Porzio*

University of Naples, Department of Mathematics and Statistics, Naples, I 80136, Italy

Sanford Weisberg†

University of Minnesota, School of Statistics, St. Paul, MN 55108, USA

Technical Report #634

March 9, 1999

Abstract

Examining lack-of-fit of a regression model is one of the fundamental problems in using regression methodology. Cook and Weisberg (1997) proposed a graphical method for this problem called *model checking plots*. The basic idea is to compare a nonparametric estimate of a mean function in the 2D scatterplot of the response versus a projection of the predictors to another estimate that is valid only if the model fit is correct. Repeating this process for many projections provides evidence concerning the lack-of-fit of the model. In this article, we provide statistics that can be used to calibrate these plots, including *reference bands* (Bowman and Young, 1996), and test statistics that are valid conditionally given the projection based on an approximating distribution (Bowman and Azzalini, 1997) or on the bootstrap (Efron and Tibshirani, 1993).

KEYWORDS: Lack-of-fit, graphical methods, regression, bootstrap, Chi-squared approximations

1 Introduction

One of the fundamental problems of regression analysis is judging the goodness of fit of a model to the data at hand. A very large literature on this problem exists. For example, in a fitted linear model with repeated observations at several

*Part of this work was performed while the first author was visiting the University of Minnesota School of Statistics under CNR Grant # 203.10.34

†Work Supported by the National Science Foundation, Grant DUE 96-52887.

values of the predictors, a standard F-test is available (e.g., Weisberg, 1985, Sec. 4.2–4.3). For binomial regression, Landwehr, Pregibon and Shoemaker (1984) proposed using cluster analysis to obtain near-replicates as the basis for lack of fit testing. Atkinson (1970) proposed tests for discriminating between separate families of models. Less formal methods of model comparison based on graphical examination of residuals (see, e.g., Cook and Weisberg, 1982, or Atkinson, 1985) are also a standard part of data analysis.

More recently, nonparametric function estimation has been used to compare models, as discussed in books by Bowman and Azzalini (1997) and Hart (1997). Other important references include Hastie and Tibshirani (1990) for generalized additive models, and Azzalini, Bowman and Härdle (1989) and Azzalini and Bowman (1993) primarily for problems with one predictor, and Young and Bowman (1995) for covariance analysis.

A new graphical approach to this old problem was presented by Cook and Weisberg (1997) using *model checking plots*. Model checking plots permit graphical assessment of a model by looking at many 2D or possibly 3D plots. In each plot, the user compares a fitted function suggested by the data to a fitted function computed from the model. If these are the same for many plots, then the model is effectively reproducing the data; otherwise, the model is inadequate.

In this article, we drive methods for calibrating model checking plots. Applying the methodology of Bowman and Young (1996), we obtain pointwise *reference bands* that can be interpreted as an acceptance region on the difference between the curves. Following Azzalini and Bowman (1993) an overall test of significance between the two fitted functions is given.

Section 2 contains a brief summary of model checking plots. Section 3 presents methodology for a reference band for equality, and contains several examples. In Section 4 the overall test is proposed and related aspects are investigated, while the last section contains further remarks.

2 Model Checking Plots

Suppose we have a regression problem in which we have observed n independent copies $(\mathbf{x}_i, y_i)$ of the random variable $(\mathbf{x}, y)$, where $\mathbf{x}$ is of dimension p and y is a scalar. The regression problem is the study of the conditional distributions $F(y|\mathbf{x})$; if F were known, then the regression problem would be completely solved. Suppose that we have a model for the regression problem, and under

the model the conditional distribution of $y|\mathbf{x}$ is given by $M(y|\mathbf{x},\theta)$, for some unknown vector of parameters θ. We assume we can obtain a consistent estimate $\hat{\theta}$ of θ, given that $M = F$, and the estimated conditional cdf is $M(y|\mathbf{x},\hat{\theta})$. The goal is to decide if $F(y|\mathbf{x}) = M(y|\mathbf{x},\theta)$.

Since both $F(y|\mathbf{x})$ and $M(y|\mathbf{x},\theta)$ are objects in $(p+1)$-dimensional space, direct graphical comparison of these functions is possible only if $p = 1$ or possibly $p = 2$. When $p > 2$, one- and two-dimensional graphs display projections of the full conditional distributions to lower dimensions, and these projections can miss relevant information. *Model checking plots* are derived from the following result: two conditional distributions $F(y|\mathbf{x})$ and $M(y|\mathbf{x})$ are the same if and only if $F(y|\mathbf{a}'\mathbf{x}) = M(y|\mathbf{a}'\mathbf{x})$ for all $\mathbf{a}, \|\mathbf{a}\| = 1$. In words, two cdfs are identical if and only if they agree on all 2D margins. This suggests that we have evidence concerning the fit of a model by examining margins. We have therefore traded the comparison of two $(p+1)$-dimensional surfaces for a potentially infinite number of comparisons of two-dimensional surfaces.

Regression problems are often summarized by the conditional mean function. Assuming the mean and the variance function are sufficiently smooth, a smoother fit to the scatterplot of y versus $\mathbf{a}'\mathbf{x}$ provides a nonparametric estimate of $\mathrm{E}_F(y|\mathbf{a}'\mathbf{x})$, the conditional mean function under the true distribution F, whether or not M is appropriate for the data. To get an estimate of $\mathrm{E}_M(y|\mathbf{a}'\mathbf{x})$, we write

$$\mathrm{E}_M(y|\mathbf{a}'\mathbf{x}) = \mathrm{E}\left[\mathrm{E}_M(y|\mathbf{x})|\mathbf{a}'\mathbf{x}\right] \approx \mathrm{E}(\hat{y}|\mathbf{a}'\mathbf{x})$$

where the equality follows from the formula for iterated conditional expectations, and the approximate equality follows if we assume that $\mathrm{E}_M(y|\mathbf{x}) \approx \hat{y}$, the fitted values under the model M. This is equivalent to assuming that $\theta \approx \hat{\theta}$, or $M(y|\mathbf{x},\theta) \approx M(y|\mathbf{x},\hat{\theta})$ (Cook and Weisberg, 1997). The effects of substituting estimates for parameters are well known in general, and are generally of lower order than the dominant terms of interest here. Consequently, we will write $M(y|\mathbf{x})$ in place of $M(y|\mathbf{x},\hat{\theta})$.

Collect the data $(\mathbf{x}_i, y_i), i = 1,\ldots,n$ into the $n \times 1$ vector $\mathbf{y}$ and the $n \times p$ matrix $\mathbf{X}$, excluding a constant column for the intercept. A direction $\mathbf{a}$ in p-dimensional space is now equivalent to the vector $\mathbf{Xa}$ in n-dimensional space. Let $\mathcal{A}$ be a set consisting of "interesting" directions $\mathbf{Xa}_1,\ldots\mathbf{Xa}_d$, where $\|\mathbf{a}_j\| = 1$. Cook and Weisberg (1997) discuss how one might choose $\mathcal{A}$ in

practice. For a fixed value of $\mathbf{a}$, the model checking plot is computed as follows:

1. Draw the scatterplot of $\mathbf{y}$ versus $\mathbf{Xa}$ and compute a nonparametric estimate of $\mathrm{E}_F(\mathbf{y}|\mathbf{Xa})$. For the results of this paper to apply, the smoother must be *linear*, which means that the smoothed values can be written as $\tilde{\mathbf{y}}_\mathbf{a} = \mathbf{W}_\mathbf{a}\mathbf{y}$ for some $n \times n$ matrix $\mathbf{W}_\mathbf{a}$ that depends on $\mathbf{Xa}$, the type of smoother, and on one or more bandwidth parameters. We have used smoothing splines (Green and Silverman, 1994), but other smoothers such as local polynomials (Bowman and Azzalini, 1997) can be used. For now, we assume that the bandwidth is fixed.

2. For fixed $\mathbf{a}$, draw a scatterplot of $\hat{\mathbf{y}}$ versus $\mathbf{Xa}$, where we have collected the fitted values $\hat{y}_i$ into $\hat{\mathbf{y}}$. The smoothed values $\tilde{\hat{\mathbf{y}}}_\mathbf{a} = \mathbf{W}_\mathbf{a}\hat{\mathbf{y}}$ provides an estimate of $\mathrm{E}_M(\mathbf{y}|\mathbf{Xa})$. If we use the same smoothing matrix $\mathbf{W}_\mathbf{a}$ for both curves, then the pointwise bias in the estimates will cancel when the two curves are compared, thus making the choice of the bandwidth somewhat less important that it is in most smoothing problems.

3. Plot the two smooths on the same graph, perhaps on the plot of $\mathbf{y}$ versus $\mathbf{Xa}$. If the model is the same as F, then the two smooths estimate the same function; if $F \neq M$, then the two smooths will be different for some choices of $\mathbf{a}$.

Model checking plots must be repeated for all $\mathbf{a} \in \mathcal{A}$. If the two curves on all model checking plots agree, then we have no evidence that the model is unacceptable.

3 Pointwise Reference Bands

To help use model checking plots, we propose adding a reference band for equality to the plot. The reference band is interpreted as a pointwise acceptance region for the hypothesis $F(\mathbf{y}|\mathbf{Xa}) = M(\mathbf{y}|\mathbf{Xa})$. It was used by Bowman and Young (1996) and by Bowman and Azzalini (1997) to compare nonparametric estimates.

Although model checking plots can be used more generally, for this article we limit ourselves to the linear regression model given by

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{X}\beta + \sigma\varepsilon \tag{1}$$

where $\theta = (\beta_0, \beta, \sigma)^T$ and ε is a random vector with zero mean and $\mathrm{Var}(\varepsilon) = \mathbf{I}$. Let $\hat{\mathbf{y}}$ be the $n \times 1$ vector of the ols fitted values under M, and let $\mathbf{H}$ the $n \times n$ projection matrix such that $\hat{\mathbf{y}} = \mathbf{Hy}$. The smoothed values are given by:

$$\tilde{\mathbf{y}}_\mathbf{a} = \tilde{\mathrm{E}}_F(\mathbf{y}|\mathbf{Xa}) = \mathbf{W}_\mathbf{a}\mathbf{y} \tag{2}$$

$$\tilde{\hat{\mathbf{y}}}_\mathbf{a} = \tilde{\mathrm{E}}_M(\mathbf{y}|\mathbf{Xa}) = \mathbf{W}_\mathbf{a}\hat{\mathbf{y}} = \mathbf{W}_\mathbf{a}\mathbf{H}\mathbf{y} \tag{3}$$

We assume that the smoothing parameter h that determines $\mathbf{W}_\mathbf{a}$ is fixed; we will address this point later.

The difference between the smoothers is

$$\tilde{\mathbf{y}}_\mathbf{a} - \tilde{\hat{\mathbf{y}}}_\mathbf{a} = \mathbf{W}_\mathbf{a}(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{W}_\mathbf{a}\mathbf{r},$$

with $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ the ols residual vector. The variance of this difference is

$$\mathbf{V}_\mathbf{a} = \mathrm{Var}_F(\tilde{\mathbf{y}}_\mathbf{a} - \tilde{\hat{\mathbf{y}}}_\mathbf{a}|\mathbf{Xa}) = \mathbf{W}_\mathbf{a}\mathrm{Var}_F(\mathbf{r}|\mathbf{Xa})\mathbf{W}_\mathbf{a}' \tag{4}$$

To estimate (4) we need an estimate of $\mathrm{Var}_F(\mathbf{r}|\mathbf{Xa})$. The simplest estimate is obtained under the assumption that $M = F$, for in this case the residuals $\mathbf{r}$ are independent of $\mathbf{X}$, and consequently

$$\mathrm{Var}_F(\mathbf{r}|\mathbf{Xa}) = \mathrm{Var}(\mathbf{r}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

We can estimate σ^2 from the fit of M using the residual mean square, $\hat{\sigma}^2 = \mathbf{r}'\mathbf{r}/(n - p - 1)$, or σ^2 could be estimated nonparametrically.

Without assuming $M = F$, write

$$\mathrm{Var}_F(\mathbf{r}|\mathbf{Xa}) = (\mathbf{I} - \mathbf{H})\mathrm{Var}_F(\mathbf{y}|\mathbf{Xa})(\mathbf{I} - \mathbf{H})$$

where $\mathrm{Var}_F(\mathbf{y}|\mathbf{Xa})$ is a diagonal matrix. If $\mathrm{Var}_F(\mathbf{y}|\mathbf{Xa}) = \tau^2\mathbf{I}$, as might be a reasonable assumption if $F \approx M$ and the predictors are approximately normally distributed, then τ^2 can be estimated from the regression of $\mathbf{y}$ on $\mathbf{Xa}$ using a nonparametric method such as the one proposed by Rice (1984). If the assumption of constant variance is not reasonable, then $\mathrm{Var}_F(\mathbf{y}|\mathbf{Xa})$ can be estimated using a variance smooth, as suggested by Ruppert, Wand, Holst, and Hössjer (1997). This last approach adds considerably to the computational complexity. Estimating variance by smoothing adds another bandwidth to the problem. In the examples discussed below, we have estimated τ^2 using Rice's

method.

The reference band for equality is obtained by superimposing a shaded band centered at the average of the two smoothers and with width given by twice the standard error of the difference between the curves at each point, in our case the square root of an estimate of the corresponding diagonal element of $\mathbf{V_a}$ given by (4) with an estimate substituted for $\mathrm{Var}_F(\mathbf{r}|\mathbf{Xa})$. Under normality of the errors this band corresponds to an asymptotic 95% pointwise confidence interval. If the fitted curves lie substantially outside the reference band, we have evidence against the hypothesis that M and F are the same.

3.1 Computations

The bounds depend on the matrix $\mathbf{W_a}$. First, suppose that the i-th row $\mathbf{w}_i$ of $\mathbf{W_a}$ were available. Let $\mathbf{QR}$ be the QR-factorization of $(\mathbf{1}, \mathbf{X})$, so $\mathbf{H} = \mathbf{QQ}'$. Using the second approximation for $\mathrm{Var}_F(\mathbf{r}|\mathbf{Xa})$, the i-th diagonal element of the matrix $\mathbf{W_a}(\mathbf{I} - \mathbf{H})\mathbf{W_a}' = \mathbf{W_a}\mathbf{W_a}' - \mathbf{W_a}\mathbf{QQ}'\mathbf{W_a}'$ can be computed as $\| \mathbf{w}_i \|^2$ - $\| \mathbf{Q}'\mathbf{w}_i \|^2$, and the estimated pointwise variance can then be computed as

$$\hat{\tau}^2(\| \mathbf{w}_i \|^2 - \| \mathbf{Q}'\mathbf{w}_i \|^2).$$

To obtain the pointwise reference band, we need the rows of the $n \times n$ smoother matrix $\mathbf{W_a}$. For graphical presentation, computing the pointwise variance is required for only a few nearly equally spaced points. Assuming that $\mathrm{E}_F(\mathbf{y}|\mathbf{Xa})$ is smooth, as few as ten to twenty evaluations of the pointwise variance give a fair representation of the bands.

Most computer programs for linear smoothers including `smooth.spline` in Splus do not provide the rows of $\mathbf{W_a}$, although providing the diagonal elements of $\mathbf{W_a}$ is common. Bowman and Azzalini (1997) provide code for local linear smoothers that returns the matrix $\mathbf{W_a}$ in full in their `sm` library. If the rows of $\mathbf{W_a}$ are not available but $\mathbf{W_a}$ is symmetric, as is the case for smoothing splines, the rows of $\mathbf{W_a}$ can be computed by repeated calls to the routine for smoothing. Let $\mathbf{u}_i$ be the $n \times 1$ vector with 1 in i-th position and zeroes elsewhere. Then the smooth of $\mathbf{u}_i$ on $\mathbf{Xa}$ will return $\mathbf{W_a}\mathbf{u}_i$, the i-th column of $\mathbf{W_a}$, equal to the i-th row by symmetry. For each evaluation of the pointwise variance, one evaluation of the smoother is required, so this method can be computationally intensive. The Bowman and Azzalini procedure substitutes computer time for computer storage, since they require an $n \times n$ array be kept. This can be prohibitive for

n of 1000 or more, but even here storage can be minimized by using binning (see, e.g., Fan and Marron, 1994).

3.2 Examples

Box and Cox (1964) presented an example that models y = number of cycles to failure for wool yarn samples in a 3^3 factorial design with factors length of test specimen, amplitude of loading cycle, and the load. We fit a first-order model to these data, and Figure 1 shows the model checking plot with $\mathbf{Xa} = \hat{\mathbf{y}}$, the fitted values. The solid line is the smoother estimated under F, while the dashed line is the smoother estimated from M. Smoothing splines were used with bandwidth chosen via cross-validation of the estimate for F, although the value of the smoothing parameter has little effect on the appearance of the plot, in agreement with results given by Bowman and Young (1996).

The reference band is the shaded area in the plot. As previously pointed out, where the smoothers lie out of the shaded area it can be inferred that the model does not describe completely the data. In Figure 1 the two fitted curves appear to disagree, and the reference band verifies this impression, since the curves are generally outside the reference band. We conclude (as did Box and Cox) that a first-order mean function does not adequately match these data.

Figure 2 shows model checking plots for the cherry tree data (Ryan, Joiner and Ryan, 1985) that were used by Bowman and Azzalini (1997, page 96) to illustrate a lack-of-fit test based on comparing two-dimensional nonparametric and parametric estimates. They used a linear model with response tree volume and two predictors tree diameter and tree height. Figure 2a shows the model checking plot for fitted values and Figure 2b for diameter. In both cases the curves are generally outside the reference bands, suggesting this model is inappropriate (dimensional considerations would of course suggest using logarithms in this problem).

As a final example, we used the fuel consumption data (Cook and Weisberg, 1994, page 207), which has response per capita motor fuel consumption and predictors per capita income, the number of vehicles per person, the tax rate, and the average miles per vehicle. The data are for the fifty U. S. states and the District of Columbia. We used a first-order linear model. In Figure 3 we show the model checking plot in the directions of the linear model fitted values and of the predictor vehicles per person. The curves appear to deviate most from the reference bands at the extreme right of the plots, where one point seems

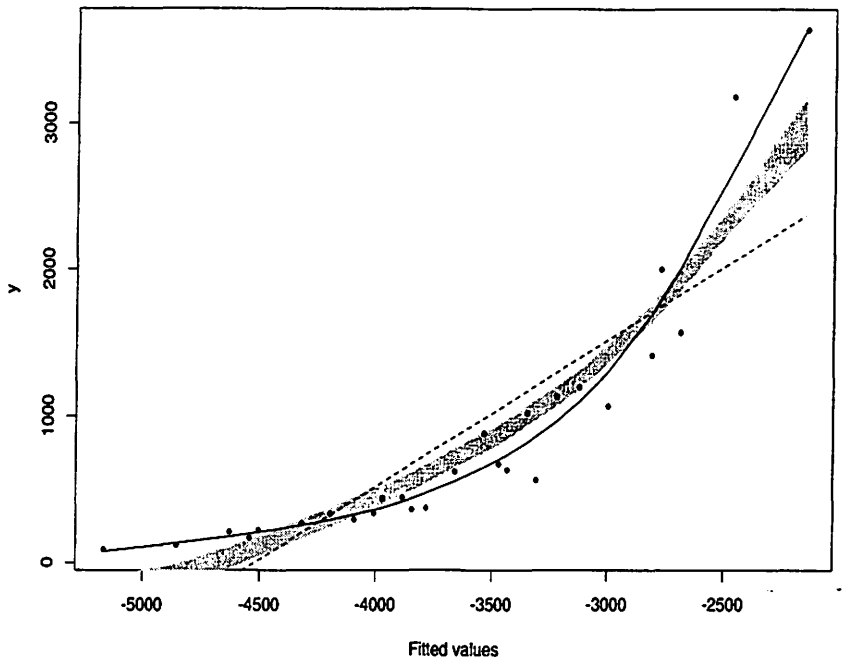


Figure 1: Model checking plots for wool data. Fitted values direction.

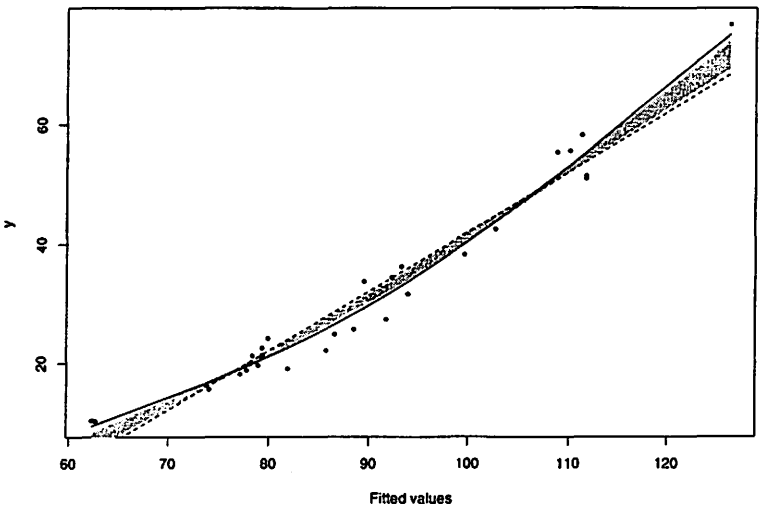
to be influential for the fit. When this case (Wyoming) is deleted, we get the model checking plots shown in Figure 4; in these two views, and any others we tried, the two curves are generally within the reference bands. In particular we note that in the direction of the fitted values the two smoothers are practically coincident and their reference band is not visible.

4 Tests

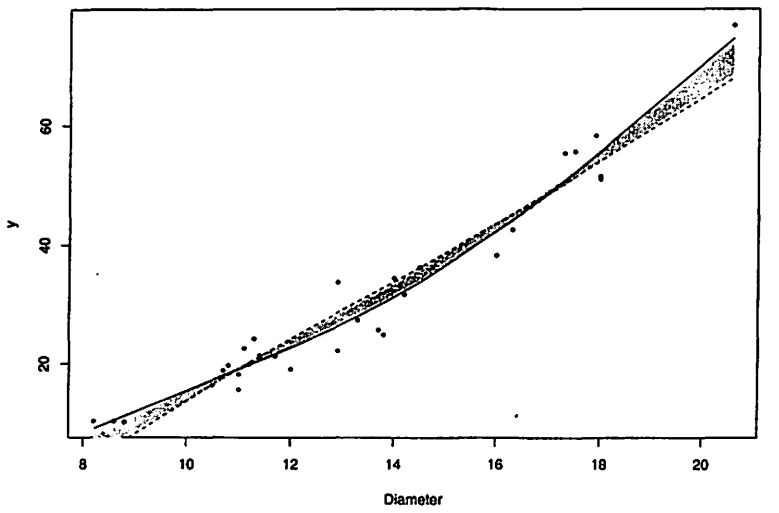
The information in a model checking plot can also be summarized with a test statistic. Comparing a nonparametric and a parametric fit with a single predictor has been described by Hart (1997) and by Bowman and Azzalini (1997), and the references they provide. We generally follow the suggestions of Bowman and Azzalini.

We consider an hypothesis test for a fixed direction **a**

$$H_0 : \mathrm{E}_F(\mathbf{y}|\mathbf{Xa}) = \mathrm{E}_M(\mathbf{y}|\mathbf{Xa}). \tag{5}$$

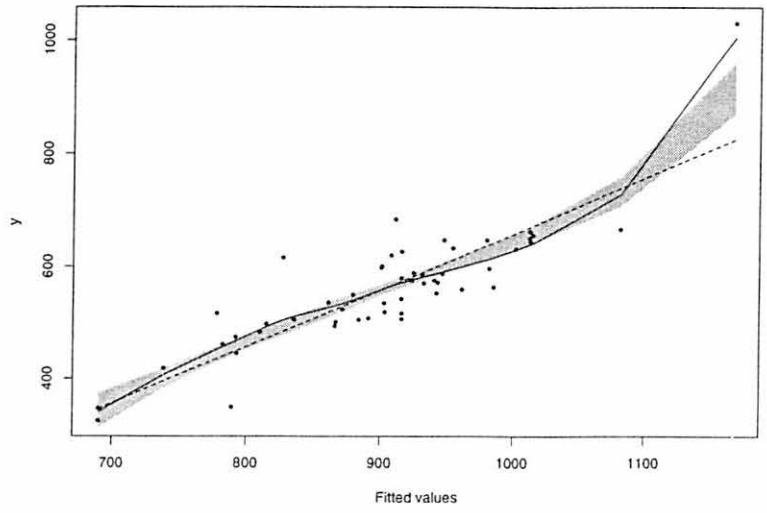


a.

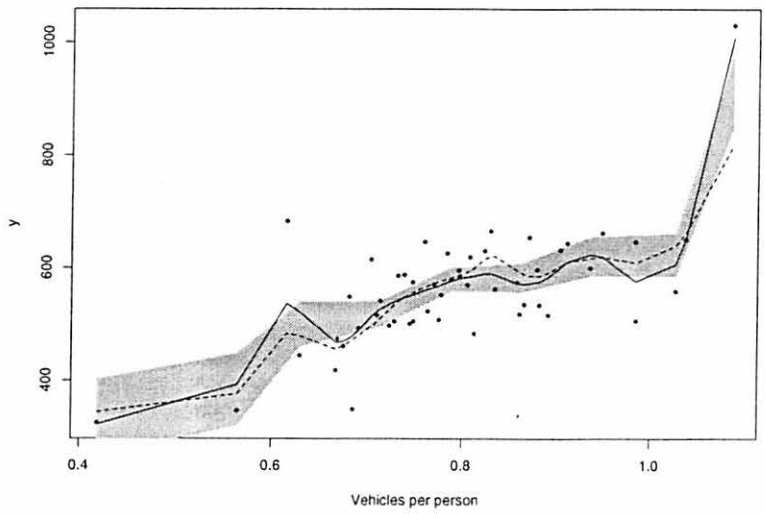


b.

Figure 2: Model checking plots for cherry trees data. Top (a): fitted values direction. Bottom (b): diameter predictor direction.

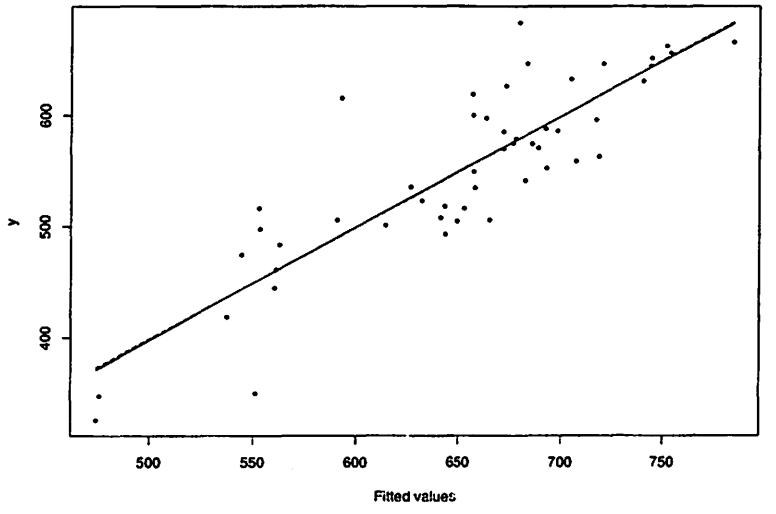


a.

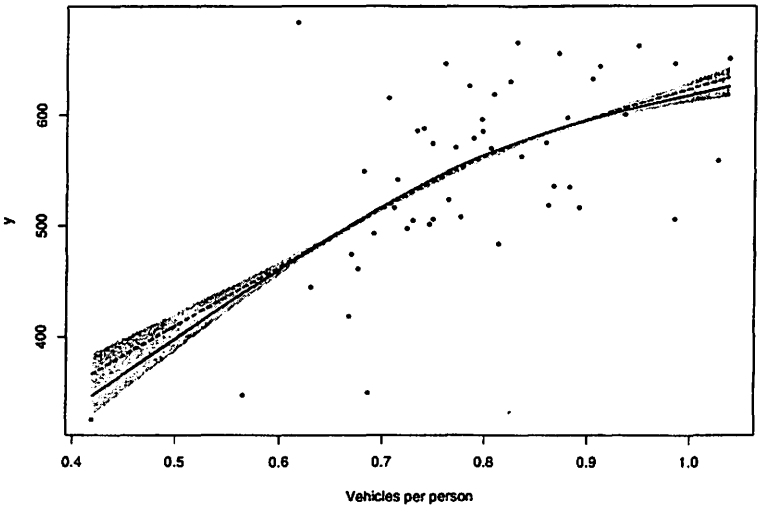


b.

Figure 3: Model checking plots for fuel data. Top (a): fitted values direction. Bottom (b): vehicles per person predictor direction.



a.



b.

Figure 4: Model checking plots for fuel data with Wyoming deleted. Top (a): fitted values direction. Bottom (b): vehicles per person predictor direction.

against a general alternative. For a test, we use estimators (2) and (3) and evaluate their difference by

$$\begin{aligned}\tilde{\mathrm{E}}_F(\mathbf{y}|\mathbf{Xa}) - \tilde{\mathrm{E}}_M(\mathbf{y}|\mathbf{Xa}) &= \tilde{\mathbf{y}}_\mathbf{a} - \tilde{\hat{\mathbf{y}}}_\mathbf{a} \\ &= \mathbf{W_a y} - \mathbf{W_a H y} = \mathbf{W_a}(\mathbf{y} - \hat{\mathbf{y}}) \\ &= \mathbf{W_a r}.\end{aligned}$$

Taking the norm of this difference and dividing by $\hat{\sigma}^2 = \mathbf{r}'\mathbf{r}/(n-p-1)$ for scale invariance, we have the test statistic

$$T_\mathbf{a} = \frac{\| \mathbf{W_a}(\mathbf{I} - \mathbf{H})\mathbf{y} \|^2}{\hat{\sigma}^2} \tag{6}$$

$$= \frac{\| \mathbf{W_a r} \|^2}{\hat{\sigma}^2} \tag{7}$$

4.1 Test Distribution

Both the numerator and denominator of (7) are quadratic forms. The numerator can be written as

$$\| \mathbf{W_a r} \|^2 = \varepsilon' \sigma (\mathbf{I} - \mathbf{H}) \mathbf{W_a}' \mathbf{W_a} (\mathbf{I} - \mathbf{H}) \sigma \varepsilon = \sigma^2 \varepsilon' \mathbf{A} \varepsilon$$

with $\mathbf{A} = (\mathbf{I} - \mathbf{H})\mathbf{W_a}'\mathbf{W_a}(\mathbf{I} - \mathbf{H})$, or $\mathbf{A} = (\mathbf{I} - \mathbf{H})\mathbf{W_a}^2(\mathbf{I} - \mathbf{H})$ if $\mathbf{W_a}$ is symmetric. If ε has a normal distribution, then using standard results on quadratic forms (e.g. Box, 1954),

$$\varepsilon' \mathbf{A} \varepsilon \sim \sum_{i=1}^{n} \lambda_i \chi_1^2. \tag{8}$$

which is a a weighted sum of χ_1^2 random variables with the λ_i equal to the eigenvalues of $\mathbf{A}$. Although the denominator is distributed as $\sigma^2 \chi^2_{(n-p-1)}$, the numerator and the denominator are not independent. Following Eagleson (1989) and Azzalini and Bowman (1993), we consider the p-value $Pr(T_\mathbf{a} > t)$ where t is the observed value of the test statistic. We can write

$$\begin{aligned}Pr(T_\mathbf{a} > t) &= Pr(\mathbf{r}'\mathbf{W_a}'\mathbf{W_a r} > t\mathbf{r}'\mathbf{r}/(n-p-1)) \\ &= Pr(\mathbf{r}'[(n-p-1)\mathbf{W_a}'\mathbf{W_a} - t\mathbf{I}]\mathbf{r} > 0) \\ &= Pr(\mathbf{r}'\mathbf{A}^*\mathbf{r} > 0). \end{aligned} \tag{9}$$

with

$$\mathbf{A}^* = (n-p-1)\mathbf{W_a}'\mathbf{W_a} - t\mathbf{I}$$

The p-value for fixed $\mathbf{a}$ can be evaluated as the probability that a quadratic form in normal variables is greater than zero. Since the quadratic form is distributed as a linear combination of χ^2 random variables, as in (8), the p-value could be computed exactly (see e.g. Khatri, 1980) but its evaluation requires computing the eigenvalues of the $n \times n$ matrix $\mathbf{A}^*$. An accurate approximation is available by matching moments.

4.2 Chi-squared Approximation

Using the method described by Buckley and Eagleson (1988) and by Azzalini and Bowman (1993), the p-value (9) can be approximated to second order by matching moments. The j-th cumulant of the quadratic form is given by

$$k_j = 2^{j-1}(j-1)!tr\left\{[\sigma^2(\mathbf{I}-\mathbf{H})\mathbf{A}^*]^j\right\}. \qquad (10)$$

Replacing σ^2 with the usual estimate $\hat{\sigma}^2$ and matching the moments of an $a\chi_b^2+c$ distribution with the moments of the quadratic form, we have

$$a = |k_3|/(4k_2), \quad b = (8k_2^3)/k_3^2, \quad c = k_1 - ab,$$

the required p-value can be approximated accurately as $1-q$, where q is the probability of lying below the point $-c/a$ in a χ^2 distribution with b degree of freedom. This procedure requires computing the diagonal elements of $(\hat{\sigma}^2(\mathbf{I}-\mathbf{H})\mathbf{A}^*)^j$, $j = 1,2,3$, still a fairly expensive calculation.

4.3 Bootstrap

When normality of errors is questionable, the bootstrap can be used to get a p-value. Using the bootstrapping residuals method (Efron and Tibshirani, 1993), resampled responses are given by $\mathbf{y}^* = \hat{\beta}_0\mathbf{1} + \mathbf{X}\hat{\beta} + \mathbf{r}^*$, with $\mathbf{r}^*$ sampled with replacement from $\mathbf{r}$, and the $\hat{\beta}$'s ols estimate under the model (1). The resampled data $(\mathbf{X}, \mathbf{y}^*)$ are then used to compute the resampled test statistic

$$T_{\mathbf{a}}^* = (n-p-1)\frac{\|\mathbf{W_a}(\mathbf{I}-\mathbf{H})\mathbf{y}^*\|^2}{\|(\mathbf{I}-\mathbf{H})\mathbf{y}^*\|^2}$$

This process is repeated B times and the bootstrap p-value is the fraction of times $T_{\mathbf{a}}^*$ exceeds the observed value of the test t.

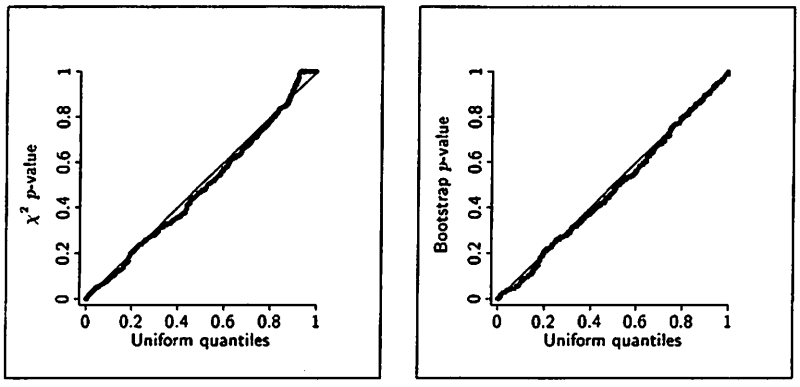
To compare the Chi-squared approximation with the bootstrap, we generated 500 samples in a simple linear regression model and computed the test and the p-values using both approximations. The samples were generated in the following way. First, a predictor $\mathbf{x} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ was generated, then the response was generated as $\mathbf{y} = \mathbf{1} + \mathbf{x} + \varepsilon$, with $\varepsilon \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$. Each sample has 100 observations and the number of bootstrap resampling for each sample was fixed at $B = 99$. In each case, the projection direction was $a = 1$; in simple regression, there is only one possible choice. The weight matrix $\mathbf{W_a}$ depends on $\mathbf{x}$ and on the bandwidth. We selected the bandwidth from the first generated sample via cross-validation, and then kept it fixed throughout the remaining simulations. Since the null hypothesis is true, the distribution of the p-values should be uniform on $(0, 1)$.

Uniform qq-plots for the p-values using the Chi-squared approximation are shown in Figure 5a, and for the bootstrap in Figure 5b. Figure 5c is a plot of the Chi-squared p-values versus the bootstrap with the 45° line superimposed. Under normality both the Chi-squared approximation and the bootstrap p-values have the expected uniform distribution. Also, the p-values computed in the two ways substantially agree; their correlation is greater than 0.988. A few trial runs have shown that if B is increased, the two p-values tends to be even closer.

4.4 Random Smoothing Matrix

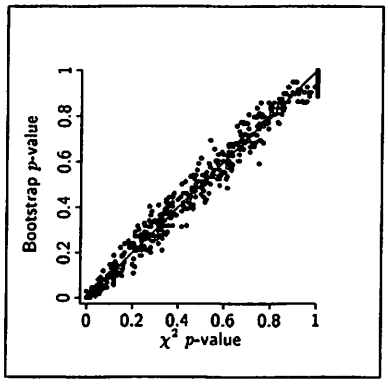
When the smoothing parameter h is a function of the data, then $\mathbf{W_a}$ is random and the Chi-square and bootstrap approximations are not appropriate. We suggest using a modified version of the bootstrap procedure. Let $\mathbf{W}_{\mathbf{a}}^*$ denote the smoothing matrix recomputed at each bootstrap resampling, using the same criterion for the choice of h used once for the test evaluation. For example, if we have used cross-validation to compute the value of the test statistic we will use cross-validation on the bootstrapped sample, and then a different h for each sample. The bootstrapped test will be given now by

$$T_{\mathbf{a}}^{**} = (n - p - 1)\frac{\| \mathbf{W}_{\mathbf{a}}^*(\mathbf{I} - \mathbf{H})\mathbf{y}^* \|^2}{\| (\mathbf{I} - \mathbf{H})\mathbf{y}^* \|^2}, \tag{11}$$



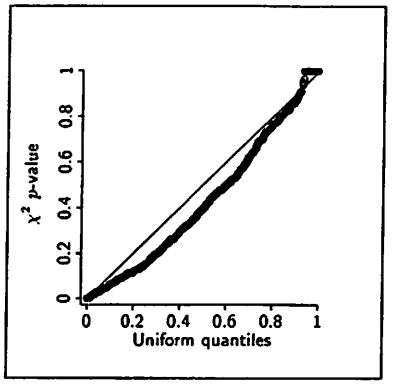
a. χ^2 p-values.

b. Bootstrap p-values.

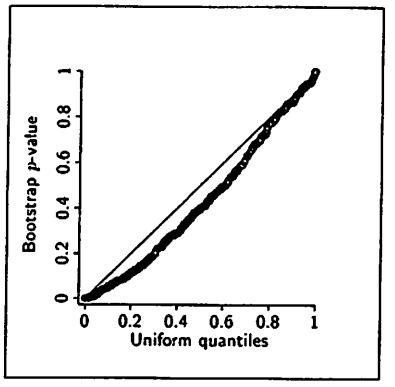


c. Comparison.

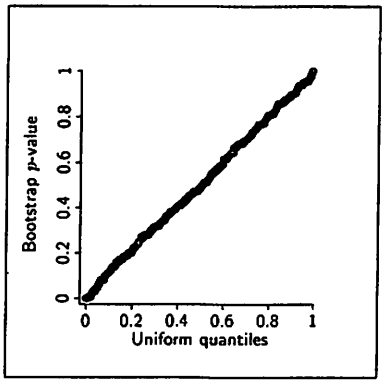
Figure 5: The p-values under fixed smoothing matrix. Uniform Q-Q plots of the χ^2 approximation (a), and using bootstrap (b), comparison of χ^2 approximation and bootstrap (c).



a. χ^2 p-values.



b. Bootstrap p-values without smoothing matrix recomputed.



c. Bootstrap p-values with smoothing matrix recomputed.

Figure 6: Uniform Q-Q plots of the p-values under random smoothing matrix. χ^2 approximation (a), bootstrap without smoothing matrix recomputed (b), bootstrap with smoothing matrix recomputed (c).

and, as before, the bootstrap p-value will be the fraction of times $T_{\mathbf{a}}^{**}$ exceeds the observed value of the test.

Figure 6 reports a simulation of this bootstrapping procedure for the setup discussed in Section 4.3 except that now we allow for random smoothing matrix computing via cross-validation a new matrix for each sample. With a random smoothing matrix, the modified bootstrap procedure reproduces the exact distribution of the test. The other two approaches yield a test p-value that is somewhat conservative, but probably sufficiently accurate for most purposes.

4.5 Power Evaluation

Additional simulations were performed to study the power of the test. The null hypothesis is given by (5) in a linear regression setting, while the alternative

hypothesis we used is

$$H_1 : E(\mathbf{y}|\mathbf{X}) = \beta_0 \mathbf{1} + (1-c)\mathbf{X}\beta + cg(\mathbf{X}) \tag{12}$$

with $0 < c \leq 1$. The model under the alternative hypothesis is composed of a linear part and a nonlinear part g. The parameter c controls the nonlinearity in the model.

The simulations were performed as follows:

1. Fix c, β_0, β and the function g. For fixed nonlinear g, $c = 0$ corresponds to the null hypothesis, and increasing c moves us away from the null hypothesis.

2. Generate an $n \times p$ predictors matrix $\mathbf{X}$, with rows independently drawn from $\mathrm{N}(\mathbf{0}, \mathbf{I}_p)$.

3. Generate a normal response variable $\mathbf{y}$ with mean given by (12) and covariance matrix $\mathbf{I}_n$.

4. Fix $\mathbf{a}$ and use the values of $(\mathbf{y}, \mathbf{X})$ to compute the weights matrix $\mathbf{W_a}$. The smoothing parameter h was computed via cross-validation, as discussed below.

5. Generate a sample as in step 3.

6. Compute the test statistic $T_\mathbf{a}$ for this sample using the weight matrix $\mathbf{W_a}$ computed in Step 4.

7. Compute the p-value of the test for each generated sample via the Chi-square approximation.

8. Repeat 5–7 N times.

The smoothing parameter h was selected in Step 4 via cross-validation either from the residual regression of $\mathbf{r}$ on $\mathbf{Xa}$ or from the response regression of $\mathbf{y}$ on $\mathbf{Xa}$. We have ignored the extra variability due to estimating h from data, resulting in conservative estimates of power. We considered two choices for the function g. First we considered the quadratic function of the linear combination β of the predictors, $g(\mathbf{X}) = (\mathbf{X}\beta)^2$, where the operation is done elementwise. According to Cook and Weisberg (in press), this is a model with one-dimensional structure because the mean function depends on the predictors only through

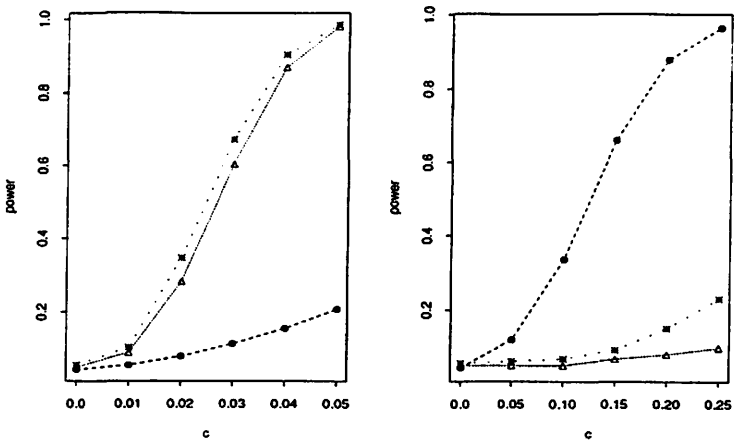


Figure 7: Estimated power curves for $p = 10$, $\alpha = 0.05$. $g(\mathbf{X}) = (\mathbf{X}\beta)^2$ left, $g(\mathbf{X}) = (\mathbf{X}\mathbf{u}_1)^2$ right. Projection directions: $\hat{\beta} = *$, $\beta = \triangle$, and $\mathbf{u}_1 = \bullet$.

a single linear combination. The second choice of g was quadratic in the first predictor, $g(\mathbf{X}) = (\mathbf{X}\mathbf{u}_1)^2$, where $\mathbf{u}_1$ is a vector of all zeroes except for the first element which is a one. For this model, the structural dimension is two. This second model is intrinsically more complicated than the first.

Although we studied a range of values for p, we only report on $p = 10$. In addition, we set $n = 100, N = 500, \beta_0 = 1$ and $\beta = \mathbf{1}$. Common random seeds were used to decrease variance of comparisons between runs.

We report only the case of $p = 10$, since results are similar for other numbers of predictors. The power depends somewhat on the way the bandwidth is estimated, with cross-validation on the response variable somewhat more powerful. Results from the simulations using cross-validation on the response are shown in Figures 7 and 8.

For both $\alpha = 0.05$ and $\alpha = 0.01$ the test is effective even for low levels of contamination (note that when $g(\mathbf{X}) = (\mathbf{X}\mathbf{u}_1)^2$ the contamination level of the data is actually $0.1c$, since only one predictor out of ten is contaminated). Power is much higher when the plotting direction $\mathbf{a}$ corresponds to the true direction of contamination than it is if $\mathbf{a}$ is incorrectly specified. This reinforces the need to generally consider many plotting directions.

4.6 Examples Revisited

The tests proposed here confirms the visual results obtained for the examples in Section 3.2. For the wool data set, the p-value is 0.005 when $\mathbf{a}$ is the fitted

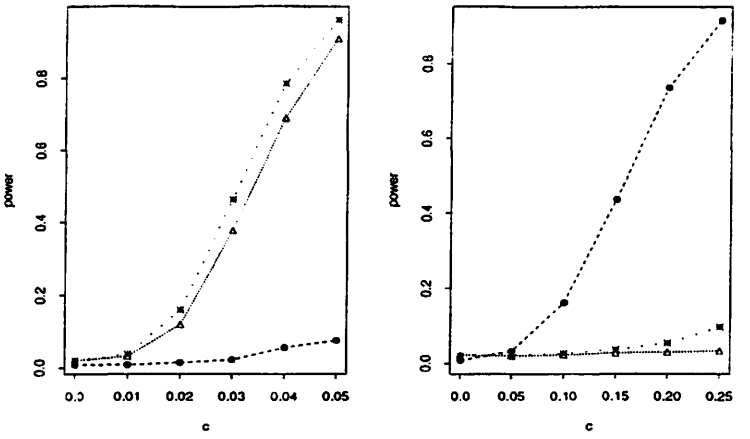


Figure 8: Estimated power curves for $p = 10$, $\alpha = 0.01$. $g(\mathbf{X}) = (\mathbf{X}\beta)^2$ left, $g(\mathbf{X}) = (\mathbf{X}\mathbf{u}_1)^2$ right. Projection directions: $\hat{\beta} = *$, $\beta = \Delta$, and $\mathbf{u}_1 = \bullet$.

value direction, confirming the graphical impression that the model with the untransformed response is not adequate. For the cherry trees data, we have had a p-value of 0.002 for both the direction of the linear model fitted values and the diameter predictor direction. For the fuel data, when Wyoming is present, the approximate p-value is 0.001 when the test is performed in the fitted values direction, and it is 0.006 in the vehicles per person predictor direction. When Wyoming is omitted ¿from the analysis, the p-values in these directions are both non significant, being respectively 0.304 and 0.321.

5 Discussion

The reference bands and the overall test are complementary and potentially useful in examining the fit or lack of fit of a regression model. As an additional point, we have found that even if the usual χ^2 approximation does not reproduce the exact distribution of quadratic forms based on linear smoothers, it yields an useful conservative bound.

Both the tests and the reference bands are conditional on the choice of the direction $\mathbf{a}$, and so a related problem is finding $\mathbf{a}$ that will maximize, at least approximately, the test $T_{\mathbf{a}}$, and then finding the distribution for the statistic found in this way. This is very similar to projection pursuit regression (Friedman and Stuetzle, 1981), and we hope to report on this related problem elsewhere.

References

[1] Atkinson, A. C. (1970). A method for discriminating between models (with discussion). *Journal of the Royal Statistical Society, series B, Methodological*, 32, 323-353.

[2] Atkinson, A. C. (1985). *Plots, Transformations and Regression*, Oxford: Oxford University Press.

[3] Azzalini, A. and Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society, series B, Methodological*, 55, 549-557.

[4] Azzalini, A., Bowman, A. and Härdle, W. (1989) On the use of of nonparametric regression for model checking. *Biometrika*, 76, 1-11.

[5] Bowman, A. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*, Oxford: Oxford University Press.

[6] Bowman A. and Young, S. (1996). Graphical comparison of nonparametric curves. *Applied Statistics*, 45, 83-98.

[7] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 25, 290-302.

[8] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, series B, Methodological*, 26, 211-246.

[9] Buckley, M.J. and Eagleson, G.K. (1988) An approximation to the distribution of quadratic forms in normal random variables. *Australian Journal of Statistics*, 30A, 150-159.

[10] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* London: Chapman and Hall.

[11] Cook, R.D. and Weisberg, S. (1994). *An Introduction to Regression Graphics.* New York: John Wiley.

[12] Cook, R.D. and Weisberg, S. (1997). Graphics for assessing the adequacy of regression model. *Journal of the American Statistical Association*, 92, 490-499.

[13] Cook, R.D. and Weisberg, S. (in press). *Applied Regression including Computing and Graphics.* New York: John Wiley.

[14] Eagleson, G.K. (1989) Curve Estimation - Whatever Happened to the Variance? *Bulletin of the International Statistical Institute*, 53, 535-551.

[15] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

[16] Fan, J. and Marron, J.S. (1994). Fast implementation of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, 3, 35-36.

[17] Friedman, J. and Stuetzle, W. (1981). Projection Pursuit Regression. *Journal of the American Statistical Association*, 76, 817-823.

[18] Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models.* London: Chapman & Hall.

[19] Hart, J. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* New York: Springer.

[20] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, London: Chapman and Hall.

[21] Khatri, C.G. (1980). Quadratic Forms in Normal Variables. *Handbook of Statistics*, ed. P.R. Krishnaiah, North-Holland Publishing Company, Vol. 1, 443-469.

[22] Landwehr, J.M., Pregibon, D. and Shoemaker, A.G. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79, 61-71.

[23] Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics*, 12, 1215-1230.

[24] Ruppert, D., Wand, M., Holst, U. and Hössjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, 39, 262–273.

[25] Ryan, B.F., Joiner, B.L. and Ryan, T.A. (1985). *Minitab Handbook*, second edition. Boston: PWS-Kent Publishing Company.

[26] Weisberg, S. (1985). *Applied Linear Regression*, second edition. New York: Wiley.

[27] Young, S.G. and Bowman, A.W. (1995). Non-parametric analysis of covariance. *Biometrics*, 51, 920-931.